Stage 6 – Data Preprocessing

Introduction

Data preprocessing is a critical step in preparing financial datasets for analysis and modeling. It ensures data quality, consistency, and usability. Poor cleaning can lead to biased results and unstable models.

Why Preprocessing Matters

Preprocessing ensures that raw financial datasets become usable for modeling. Every decision—whether to drop, fill, or scale data—encodes assumptions. Poor cleaning leads to biased results and model instability.

Missing Data

Types:

- MCAR: Missing completely at random
- MAR: Missing at random, depends on observed data
- MNAR: Missing not at random, depends on unobserved data

Strategies:

- Drop rows or columns with missing values
- Fill missing values with mean, median, or mode
- Forward/backward fill
- Custom imputation based on domain knowledge

Visualizations:

- Heatmaps
- Missingno matrices

Impact: Poor handling of missing data can introduce bias or errors in modeling.

Filtering

- Remove invalid entries, e.g., negative prices
- Drop rows/columns with excessive missingness

Normalization & Scaling

- MinMaxScaler: rescales features to [0,1]
- StandardScaler: rescales features to mean=0, std=1
- Important for algorithms sensitive to feature magnitude
- Visual comparisons (e.g., histograms) help verify scaling

Column Type Corrections

- Numeric conversions: remove symbols, convert string→float

- Date parsing: string→datetime
- Categoricals: string→category
- Ensures computations and modeling behave correctly

Documentation & Reproducibility

- Record assumptions and chosen strategies
- Use modular functions to support code reuse
- Validate cleaned datasets to prevent silent errors
- Save processed datasets for reproducibility

Practical Example

Raw stock dataset:

- Missing volumes
- Negative price entries
- Features with different magnitudes

Steps:

1. Fill missing volumes with median or mean
2. Drop negative price rows
3. Scale features with StandardScaler or MinMaxScaler
4. Save the dataset and document assumptions

Key Takeaways

- Cleaning is foundational, not optional
- Each preprocessing choice reflects assumptions
- Documentation and reproducible workflows are critical
- Missing data handling, filtering, scaling, and type corrections are essential skills

Preprocessing Assumptions to Communicate to Stakeholders

When cleaning and preprocessing data, each choice encodes assumptions. It is important to document these for transparency, reproducibility, and stakeholder understanding.

1. Missing Data Handling

- Filling missing numeric values with median assumes missingness is MCAR or MAR (not systematically biased).
- Forward/backward fill assumes temporal continuity in time series data.
- Dropping rows assumes missing rows are not critical to analysis.
- Imputation affects averages, distributions, and model training.

2. Filtering / Data Cleaning

- Removing negative or out-of-range values assumes they are errors or invalid entries.
- Dropping columns or rows with excessive missingness assumes those data are non-essential.

- Rare but valid events might be lost if thresholds are too strict.

3. Scaling / Normalization

- StandardScaler assumes features are roughly normally distributed.
- MinMaxScaler assumes min and max values are representative and not extreme outliers.
- Scaling changes interpretation of magnitudes; coefficients or distances may be affected.

4. Column Type Corrections

- Converting strings to numeric assumes no hidden characters or formatting issues.
- Parsing dates assumes consistent date format.
- Categoricals assume a finite, discrete set of values.
- Wrong types can break computations or modeling.

5. Reproducibility & Modularity

- Using modular functions assumes future datasets follow similar structure and patterns.
- Documenting assumptions ensures that preprocessing is transparent and results are interpretable.

> **Tip:** Always communicate these assumptions to stakeholders, so they understand the limitations and decisions made during preprocessing.